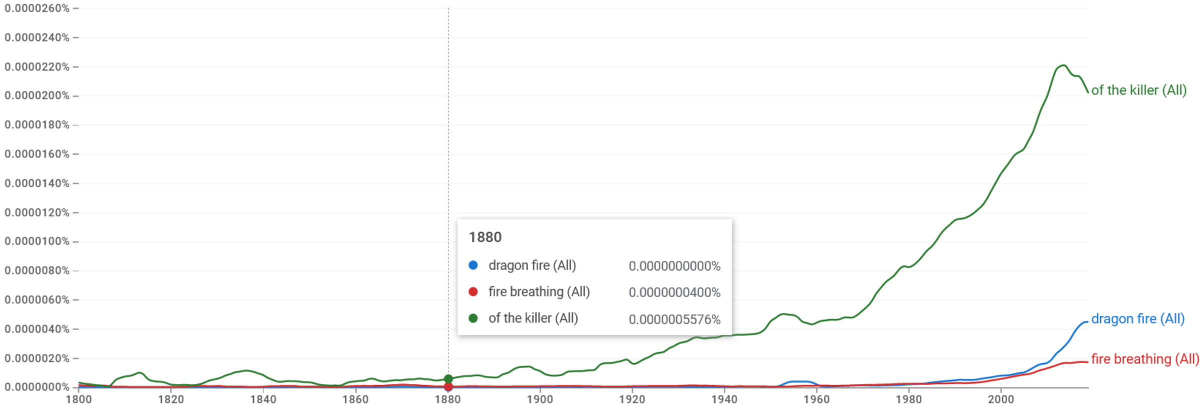
N-grams consist of groups of n adjacent words in a sentence. One word n-grams, unigrams, are just tokens. Two word n-grams, bigrams allow for more interesting uses, like calculating the chance a specific word is next to another specific word. Even larger n-grams can be used for more complicated cases, although, them being larger makes them more complicated to process. In language models, n-grams are useful for finding a word’s relation to other words via adjacency, and larger n-grams may even find the relation of a group of words to another group of words.

N-grams can be used to guess which part of speech an unknown word is by checking adjacent words’ part of speech and location to said word. They can be used to generate sentences by selecting a token that will likely come after the last word. (However, this won’t generate any new words, may cause long sentences, and be grammatically incorrect.) N-grams can also be used to compare styles of writers, seeing how words are used differently by inspecting different contexts writers use them.

Probabilities for unigrams and bigrams are calculated from how many times they occur in the source text divided by the total number of them existing in said text. Since probabilities are based off the source text, it is very important to have a source text that is large, accurate for the scenario, and shows as many variations as it can for the scenario. Of course, in the case that a word is not in the source text, it would return zero, which makes the whole text analyzed impossible according to the source. For more accurate results smoothing is applied which ensures that even for new words, there is a probability. An example of smoothing for bigrams is Laplace smoothing which is having bigram occurrence plus one divided by unigram occurrence of the first letter plus the total vocabulary. For a new word this would be one divided by the total vocabulary.

Language models can be evaluated by their vocabulary, accuracy, and size. A good example of all of these is Google's n-gram viewer which has a large source text, knows text occurrence through the ages, and other things. Below is an example of n-gram use which shows the occurrence of a few example n-grams.


dragon fire (All) 0.0000000000%
fire breathing (All) 0.0000000400%
of the killer (All) 0.0000005576%
of the killer (All)
dragon fire (All)
fire breathing (All)